Flexible Smoothing via the Locally Adaptive COmponent Selection and Shrinkage Operator (LACOSSO).

Alvaro Nosedal-Sanchez[a], Curtis B. Storlie[b]

[a] Indiana University of Pennsylvania
[b] Los Alamos National Laboratory

Date: November 2, 2011

Abstract

A new method for nonparametric function estimation is proposed, which allows for a more flexible estimation of the function in regions of the domain where it has more curvature. The Locally Adaptive COmponent Selection and Shrinkage Operator (LACOSSO) is a method for spatially adaptive nonparametric regression. This method is derived using a theoretical framework provided by reproducing kernel Hilbert spaces. In this framework, knowledge of the reproducing kernel of the functional space in question is essential. The reproducing kernel is derived with an intuitive approach. A theorem that establishes the optimal MSE convergence rate of the method and the conditions needed to achieve this convergence rate is also presented. In depth simulation studies demonstrate LACOSSO's performance is typically better, and at worst comparable, to the performance shown by its competitors.

1 Introduction

Nonparametric Regression has proven to be a very useful methodology, with applications to a large list of modern problems such as computer models, image data, environmental processes, to name a few. The nonparametric regression model is given by

$$y_i = f_0(x_i) + \epsilon_i,\ i = 1, 2, ..., n \tag{1}$$

where f_0 is an unknown regression function and ϵ_i are independent error terms.

Smoothing splines are among the most popular methods for estimation of f_0 due to

their good empirical performance and sound theoretical support (Cox 1983, Speckman 1985, Eubank 1999, van de Geer 2000). It is usually assumed without loss of generality that the domain of f_0 is [0,1]. Let $f^{(m)}$ denote the m^{th} derivative of f. The smoothing spline estimate $\hat{f}$ is the unique minimizer of

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 \left(f^{(m)}(x)\right)^2 dx \tag{2}$$

over all functions, f, in m^{th} order Sobolev space,

$$S^m = \{f : f^{(j)} \text{ is absolutely continuous for } j = 1, ..., m-1 \text{ and } f^{(m)} \in L_2\}$$

The minimizer of (2) trades off fidelity to the data (in terms of residual sum of squares) against smoothness of the reconstructed curve (in terms of the integrated squared derivative of order m, where m is typically taken to be two). The smoothing spline solution uses a global smoothing parameter λ which implies that the true underlying mean process has a constant degree of smoothness. In this paper a more general, "spatially adaptive", framework is investigated, that accommodates varying degrees of smoothness by seeking solutions where the smoothness penalty depends on the region of the domain where the function estimation is occurring. The estimator is obtained within a Reproducing Kernel Hilbert Space framework (Wahba 1990, Gu 2002).

There are many approaches to surface fitting using spatially adaptive knot placement (basis function selection) with regression splines; see Friedman & Silverman (1989), Stone, Hansen, Kooperberg & Truong (1997), Luo & Wahba (1997), and Hansen & Kooperberg (2002). However, the properties of these estimators are difficult to study analytically since they are the result of an algorithm and not an explicit solution to an optimization problem. Pintore, Speckman & Holmes (2006) use a piecewise constant function for λ in (2). However, this form of $\lambda(x)$ requires specifying the number of knots, the knot locations, and the values of $\lambda(x)$ in between knot locations. This was accomplished

by selecting one of several candidate knot location options and λ values between the knots via GCV. Unfortunately this leads to a smoothing method with large number of smoothing parameters whose values need to be selected. The Loco-Spline procedure of Storlie, Bondell & Reich (2010) uses a spatially varying penalty based on an initial estimate. The final estimate is penalized less where the initial estimate indicates more curvature is needed. However, this procedure can be unstable for small sample sizes and is computationally expensive for larger samples.

Here, a method is proposed which breaks down the interval $[0, 1]$ into p disjoint sub-intervals. Then p functional components in $[0, 1]$ are defined, which have two important features. First, the purpose of each of these p components is to estimate the true function locally, i.e., in only one of the sub-intervals. Second, even though all components are defined on the entire domain, i.e., $[0, 1]$, a component has curvature only in one of the afore mentioned intervals. The p local estimates are then added together to produce a function estimate over the entire $[0, 1]$ interval. This is similar in spirit to the method of Pintore et al. (2006). However, in the proposed method, the additional flexibility that comes from finding these p local functional estimates does not come at any additional computational cost. In spite of having p components there is no need to specify (e.g., choose via cross validation) p smoothing parameters. Theory from COmponent Selection and Shrinkage Operator (COSSO) (Lin & Zhang 2006), reduces the problem of specifying these p smoothing parameters to specifying only one smoothing parameter without a loss in flexibility. In fact, empirical studies indicate superior performance of COSSO in the additive model framework over that for the traditional additive model (Hastie & Tibshirani 1990), see Storlie, Bondell, Reich & Zhang (2011), for example. For the same reason (i.e., one tuning parameter as opposed to many), the empirical studies in this paper indicate superior performance of the proposed method to that suggested in Pintore et al. (2006).

Section 2 provides a review the COSSO framework that will be used to solve for

the proposed estimator. In Section 3, the Locally Adaptive COmponent Selection and Shrinkage Operator (LACOSSO), a new method for spatially adaptive nonparametric regression, is presented. Section 4 is devoted to computational details, e.g., finding the reproducing kernel of the functional spaces needed to solve the proposed optimization problem. In Section 5 a theorem that establishes the optimal MSE convergence rate of LACOSSO is presented. The proof of this result can be found in the Supplementary Material. Section 6 presents results from a simulation study and an example dataset to compare LACOSSO to other existing methods. In all the examples presented, LACOSSO's performance is better, or comparable, to the performance shown by its competitors. Section 7 concludes the chapter with some closing remarks and mentions areas worthy of future exploration.

2 Smoothing Spline ANOVA and COSSO

In this section only the necessary concepts of Smoothing Spline (SS)-ANOVA needed for the development of LACOSSO are reviewed. For a more detailed overview of Smoothing Splines and SS-ANOVA see Wahba (1990), Wahba, Wang, Gu, Klein & Klein (1995), Schimek (2000), Gu (2002), and Berlinet & Thomas-Agnan (2004). For a gentle introduction to RKHS and penalized regression, see Nosedal-Sanchez, Storlie, Lee & Christensen (2011).

In the smoothing spline literature it is typically assumed that $f \in F$ where F is a reproducing kernel Hilbert space (RKHS). Denote the reproducing kernel (r.k.), inner product, and norm of F as K_F, $\langle \cdot, \cdot \rangle_F$, and $|| \cdot ||_F$ respectively. Often F is chosen to contain only functions with a certain degree of smoothness. For example, functions on one variable are often assumed to belong to the second order Sobolev space, $S^2 = \{f : f, f'$ are absolutely continuous and $f'' \in L^2[0,1]\}$.

A RKHS F can always be written as

$$F = \{1\} \oplus \{\bigoplus_{j=1}^{p} F_j\}, \tag{3}$$

where $\oplus$ represents the direct sum operation, $F_1, ..., F_p$ is some orthogonal decomposition of the space, and each of the F_j is itself a RKHS. A familiar example of such a decomposition is the additive model $f(\boldsymbol{x}) = b_0 + \sum_{j=i}^{p} f_j(x_j)$ when there is more than one predictor.

A traditional smoothing spline type method finds $\hat{f} \in F$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} \theta_j^{-1} ||P^j f||^2 \tag{4}$$

where $P^j f$ is the orthogonal projection of f onto F_j and $\theta \geq 0$. If $\theta_j = 0$, then the minimizer is taken to satisfy $||P^j f||^2 = 0$. We use the convention $0/0 = 0$ throughout this paper. The smoothing parameter λ is confounded with the $\theta' s$, but is usually included in the setup for computational purposes.

The COSSO (Lin & Zhang 2006) penalizes on the sum of the norms instead of the squared norms as in the traditional smoothing spline and hence achieves sparse solutions (e.g., some of the functional components are estimated to be exactly zero). Specifically, the COSSO estimate, $\hat{f}$, is given by the function $\hat{f} \in F$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} ||P^j f||_F \tag{5}$$

where λ is a smoothing parameter.

The Adaptive COSSO (ACOSSO) improves upon COSSO by using individually weighted norms to smooth each of the components. Specifically, ACOSSO selects as the estimate

the function $f \in F$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} w_j \|P^j f\|_F \tag{6}$$

where $0 < w_j < \infty$ are weights that can depend on an initial estimate of f which we denote $\tilde{f}$. The w_js are not tuning parameters in the sense that they would need to be chosen by cross validation. For more details about ACOSSO see Storlie et al. (2011).

Finally, it is possible to give an equivalent form of (6) that is useful for computational purposes. Consider the problem of finding $[\theta_1, ..., \theta_p] \in \Re^p$ and $f \in F$ to minimize

$$\min_{f \in F} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \lambda_0 \sum_{j=1}^{p} \frac{w_j^2}{\theta_j}\|P^j f\|_F^2 + \lambda_1 \sum_{j=1}^{p}\theta_j \tag{7}$$

subject to $\theta_j \geq 0$, $j = 1, ..., p$, λ_0 is a constant that can be fixed at any positive value, and λ_1 is a smoothing parameter. For a given λ in (6), there is a value of λ_1 in (7) that will result in the same minimizing function $\hat{f}$. See Storlie et al. (2011) for a proof of this equivalence. Knowledge of which value of λ_1 corresponds to which value of λ is typically not needed, since the smoothing parameter is usually not pre-specified, rather it is chosen based on some goodness of fit measure anyhow. The minimization in (7) has the same flexibility as a minimization with p smoothing parameters $\theta_1, \ldots, \theta_p$. However, the θ_j are treated as if they are additional model parameters, then they are also penalized (in the last term). This is similar to modeling the θ_j with a hyper-prior in a hierarchical Bayesian framework.

3 A locally Adaptive Estimator

The penalty term on the right of (2) is an overall measure of the roughness of the function over the domain. The tuning parameter λ controls the trade-off in the resulting estimate

between smoothness and fidelity to the data; large values of λ will result in smoother functions while smaller values of λ result in rougher functions but with better agreement to the data. In many cases the underlying function changes more abruptly in some regions than in others. In situations like this the global penalty will cause the smoothing spline estimator to either over-smooth in some regions and/or under-smooth in others (Storlie et al. 2010).

Consider spatially adaptive estimators which are defined by the explicit function minimization problem,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \lambda \sum_{j=1}^{p} w_j \left\{ \int_{\tau_{j-1}}^{\tau_j} [f''(x)]^2 dx \right\}^{1/2} \tag{8}$$

over all functions, $f \in S^2$, for given knots $0 = \tau_0 < \tau_1 < \cdots < \tau_p = 1$. The knots need to be pre-specified (they could be chosen to be equally spaced on the quantiles of x, for example).

An equivalent minimization to (8) which is more convenient for computational purposes is

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i - \sum_{j=1}^{p} f_j(x_i)\right)^2 + \lambda \sum_{j=1}^{p} w_j \left\{ \int_{\tau_{j-1}}^{\tau_j} [f_j''(x)]^2 dx \right\}^{1/2} \tag{9}$$

over $b_0, b_1 \in \Re$, and all functions $f_1 \in S_1^*, \ldots, f_p \in S_p^*$, where

$$S_j^* = \{f : f \text{ and } f' \text{ absolutely continuous, } f'' \in L_2 \text{ , with } f(x) = 0 \text{ if } x \in [0, \tau_{j-1}),$$

f is linear for $x \in (\tau_j, 1]\}$ $j = 1, \ldots, p$.

The proposed estimator in (8) has several important properties. First, this formulation allows for the functional estimate to vary adaptively with x allowing for more/less penalty in regions of the domain where it is beneficial. This is accomplished by breaking

the function down into the p functional components. Second, the estimator is not penalizing the squared norm, rather the norm of each of these p variables, as in the COSSO framework. In doing so the estimator also inherits the computational advantages from COSSO as discussed further in Section 4. Third, there are p new elements in the minimization problem, $w_1, w_2, ..., w_p$. However, these are not smoothing parameters (i.e., they will not be estimated), rather they are weights that can depend on an initial estimate of f which we denote $\tilde{f}$. For example, f could initially be estimate via the traditional smoothing spline (a particular way of finding these quantities will be discussed in the next subsection). Finally, there is only one smoothing parameter to choose via cross validation or similar means which keeps computation more feasible, and results in better performance in practice.

3.1 Specifying w_j.

Given an initial estimate $\tilde{f}$, we wish to construct w_j's so that the prominent functional components enjoy the benefit of a smaller penalty relative to less important functional components. In contrast to the adaptive LASSO procedure for linear models (Zou 2006), there is no single coefficient, or set of coefficients, to measure importance of a variable. One possible scheme would be to make use of an estimate of the RKHS norm used in the COSSO-like penalty and set

$$w_j = \|\tilde{f}_j\|_F^{-\gamma}. \tag{10}$$

We suggest the following procedure to specify the w_js:

1. Set $w_j = 1$ for $j = 1, 2, ..., p$ in (8). By doing this, the same importance is placed on each of the functional components. With this choice of w_j's (8) becomes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\sum_{j=1}^{p}\|P^j f\|_F. \tag{11}$$

The solution to the above minimization problem gives us $\tilde{f}$.

2. Set $w_j = \|P^j \tilde{f}\|_F^{-\gamma}$, for some parameter γ. We have found that setting $\gamma = 1$ or $\gamma = 2$ provides good results in practice.

4 Computation

4.1 Solving LACOSSO with the RKHS framework

If each of the S_j^* is endowed with the inner product

$$\langle f, g \rangle = \int_0^1 f''(x) g''(x) dx$$

then each of the S_j^* are orthogonal in the space $F = \bigoplus_j S_j^*$. It then becomes clear that (9) is a special case of (6). Thus, using the equivalence of (6) and (7), the minimization in (9) can be written as the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i - \sum_{j=1}^{p} f_j(x_i) \right)^2 + \lambda_0 \sum_{j=1}^{p} \frac{w_j^2}{\theta_j} \left\{ \int_{\tau_{j-1}}^{\tau_j} [f_j''(x)]^2 dx \right\} + \lambda_1 \sum_{j=1}^{p} \theta_j \quad (12)$$

over $b_0, b_1 \in \Re$, and all functions $f_1 \in S_1^*, \ldots, f_p \in S_p^*$.

The algorithm used to solve (7) is detailed in Storlie et al. (2011) and can be used here to solve (12) as well. It hinges on the representation theorem of Wahba (1990) to write the solution to (12) in the form

$$\hat{f}(x) = \hat{b}_0 + \hat{b}_1 + \sum_{i=1}^{n} \hat{c}_i \sum_{j=1}^{p} R_j(x_i, x) \quad (13)$$

The main ingredient needed in the ACOSSO algorithm is thus the reproducing kernel R_j for each orthogonal subspace S_j^*. These reproducing kernels are presented in the next subsection. But first, a simple form for the w_j in (10) is presented based on the initial

estimate $\tilde{f}$. Write the initial estimate in the form $\tilde{f} = \tilde{b}_0 + \tilde{b}_1 + \sum_{i=1}^{n} \tilde{c}_i R_j(x_i, x)$ as given in (13). We know that $\tilde{f}_j$ is given by the orthogonal projection of $\tilde{f}$ onto H_j, which is $P^j \tilde{f} = \sum_{i=1}^{n} \tilde{c}_i R_j(x_i, \cdot)$. Hence,

$$\|P^j \tilde{f}\|_F^2 = \langle \sum_{i=1}^{n} \tilde{c}_i R_j(x_i, \cdot), \sum_{i=1}^{n} \tilde{c}_i R_j(x_i, \cdot) \rangle = \tilde{\boldsymbol{c}}' \boldsymbol{\Sigma} \tilde{\boldsymbol{c}}, \tag{14}$$

or

$$w_j = (\tilde{\boldsymbol{c}}' \boldsymbol{\Sigma} \tilde{\boldsymbol{c}})^{-\gamma/2} \tag{15}$$

4.2 Finding the Reproducing Kernels

Finding the reproducing kernel (r.k.) directly for the S_j^* would be difficult. Hence, we instead make use of the connection between a RKHS H with reproducing kernel $R(s,t)$ and a Gaussian Process (GP) with covariance $K(s,t) = R(s,t)$. The connection is based on the following result, let $\{X(t), t \in T\}$ be a real Gaussian Stochastic Process defined on a probability space, with mean function $E[X(t)] = 0$ and covariance $K(s,t) = E[X(t)X(s)]$. It is well known, see Parzen (1961), that K determines a Hilbert space $H(K)$, called the RKHS of K, which has the following properties: $K(\cdot, t) \in H(K)$ and $\langle f, K(\cdot, t) \rangle = f(t)$ for every $t \in T$. We say that such an $\{f(t), t \in T\}$ is a representation of the process $\{X(t), t \in T\}$. Before finding the r.k. for the space of functions S_j^*, we first present an example that will provide some intuition into the search process.

Example (Integrated Brownian Motion). Denote by H the collection of functions f with $f'' \in L^2[0,1]$ and consider the subspace $W_2 = \{f(x) \in H : f, f' \text{ absolutely continuous and } f(0) = f'(0) = 0\}$. Define the inner product on H as

$$\langle f, g \rangle = \int_0^1 f''(t) g''(t) dt \tag{16}$$

It can be shown (see Nosedal-Sanchez et al. (2011) for example), that

$$R(s,t) = \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6} \tag{17}$$

Now, also consider a stochastic process with $K(s,t) = R(s,t)$. Let $\{X(t), t \in [0,1]\}$ be the Wiener process. Define a new stochastic process $\{Z(t), t \in [0,1]\}$ by

$$Z(t) = \int_0^t X(s)ds \tag{18}$$

The process $\{Z(t), t \in [0,1]\}$ the integrated Wiener process or integrated Brownian process. It can be shown (Parzen 1962) that $E[Z(t)] = 0$ and

$$E[Z(s)Z(t)] = \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6}. \tag{19}$$

Thus W_2 is a representation of $\{X(t), t \in [0,1]\}$.

Using intuition gained from this example, the reproducing kernels for the S_j^* can be derived. The steps needed to find the r.k. are the following:

(i). Use intuition to guess at the G.P. $\{X(t), t \in [0,1]\}$ that corresponds to the RKHS H for which the form of the r.k. is desired R.

(ii). Find the covariance function K for $X(t)$.

(iii). Demonstrate that $R = K$ is such that $R(\cdot, t) \in H$ and $\langle R(\cdot, t), f\rangle = f(t)$ for $f \in H$, so that R is the unique r.k. for H.

For ease of presentation, first consider the simplest case: two subintervals. Let $\tau_1 \in [0,1]$, given τ_1 break down the $[0,1]$ interval into two subintervals. The basic idea is to express the function $f(x)$ as

$$f(x) = \alpha + \beta x + f_1(x) + f_2(x) \tag{20}$$

where $f_1 \in S_1^*$ and $f_2 \in S_2^*$. Equation (20) expresses $f(x)$ as a function in the space $F = \{1\} \oplus \{x\} \oplus S_1^* \oplus S_2^*$. To apply the RKHS framework and computational solution of ACOSSO to this problem the RKHS's $H_1 \subset S_1^*$ and $H_2 \subset S_2^*$ need to be defined and the corresponding r.k.'s R_1 and R_2, respectively, need to be derived. It is not necessary to define a RKHS for the constant or linear term since they lie in the null space of the penalty in (9).

Define $H_1 = S_1^*$ with inner product $\langle f_1, g_1 \rangle_{H_1} = \int_0^1 f_1''(t) g_1''(t) dt$. Similarly $H_2 = S_2^*$ with inner product $\langle f_2, g_2 \rangle_{H_2} = \int_0^1 f_2''(t) g_2''(t) dt$. As previously mentioned, the locally adaptive estimator in (9) now becomes a special case of ACOSSO in (6).

First, we find the r.k. for H_1. By the definition of H_1, $f_1 \in H_1$ implies f_1 has curvature only in $[0, \tau_1]$, $f_1 \in S^2$ (where S^2 represents 2nd order Sobolev Space) and the inner product for H_1 is the same as that in the previous example involving W_2 and integrated Brownian motion.

In an effort to find the GP representation of H_1 construct a Gaussian process as follows

$$Z_1(t) = \begin{cases} \int_0^t X(s) ds & 0 \le t \le \tau_1 \\ \int_0^t X(s) ds + X(\tau_1)(t - \tau_1) & \tau_1 \le t \le 1 \end{cases}$$

where $X(t)$ is a Wiener Process or Brownian motion.

The covariance function for Z_1, K_1, is a function whose domain is $\Re \bigotimes \Re$. Given τ_1, a couple (s, t) can fall into one of four regions: (i) $s \in [0, \tau_1]$ and $t \in [0, \tau_1]$, (ii) $s \in (\tau_1, 1]$ and $t \in (\tau_1, 1]$, (iii) $s \in [0, \tau_1]$ and $t \in (\tau_1, 1]$ and (iv) $t \in [0, \tau_1]$ and $s \in (\tau_1, 1]$.
Below, $K_1(s, t)$ is defined for each of these cases. However, knowing that $K_1(s, t)$ must be a symmetric function cases (iii) and (iv) are the same so there are really only three cases. The calculations for each case are carried out in Section A.1 of the Supplementary

Material, but the results are given here for convenience. $K_1(s,t)$ is defined as follows

$$= \begin{cases} \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6} & \text{for } s,t \in [0,\tau_1] \\ \frac{\tau_1^3}{3} + \frac{(\max(s,t)-\tau_1)\tau_1^2}{2} + \frac{(\min(s,t)-\tau_1)\tau_1^2}{2} + \frac{2[\min(s,t)-\tau_1][\max(s,t)-\tau_1]\tau_1}{2} & \text{for } s,t \in (\tau_1,1] \\ \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6} & \text{otherwise} \end{cases} \tag{21}$$

In Section A.2 of the Supplementary Material, it is demonstrated that $K_1(s,t) = R_1(s,t)$ has the reproducing property, and hence, is the r.k. for H_1. The r.k. $K_1(s,t)$ also depends on τ_1. For ease of notation, make this dependence explicit by writing $K^*(s,t,\tau_1) = K_1(s,t)$.

Now, by the definition of S_2^*, the functions in S_2^* are functions equal to zero in $[0,\tau_1]$ and with curvature in $(\tau_1,1]$. Parallel to that above, define the stochastic process $Z_2(t)$ as follows

$$Z_2(t) = \begin{cases} 0 & 0 \le t \le \tau_1 \\ \int_{\tau_1}^t X(s-\tau_1)ds & \tau_1 \le t \le 1 \end{cases}$$

where $X(t)$ is a Wiener Process or Brownian motion. From the above definition, it should be clear that $Z_2(t)$ is a shifted version of $Z_1(t)$, taking on a value of exactly 0 in the region where Z_1 is nonlinear, and providing nonlinearity in the region where Z_1 is linear.

The covariance function of Z_2 is

$$K_2(s,t) = \begin{cases} K^*(\frac{s-\tau_1}{1-\tau_1}, \frac{t-\tau_1}{1-\tau_1}, \frac{\tau_2-\tau_1}{1-\tau_1}) & \text{for } s,t \in (\tau_1,1] \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Again, it is demonstrated in Section A.2 of the Supplementary Material that $K_2(s,t) = R_2(s,t)$ has the reproducing property, and hence, is the r.k. for H_2.

One can do something analogous to derive the reproducing kernels in the case of multiple knots $0 = \tau_0 < \tau_1 < \cdots < \tau_{p-1} < \tau_p = 1$. In the general case that $H_j = S_j^*$ with

inner product $\langle f_j, g_j \rangle_{H_j} = \int_0^1 f_j''(t) g_j''(t) dt$, the r.k. for H_j is

$$K_j(s,t) = \begin{cases} K^*(\frac{s-\tau_{j-1}}{1-\tau_{j-1}}, \frac{t-\tau_{j-1}}{1-\tau_{j-1}}, \frac{\tau_j-\tau_{j-1}}{1-\tau_{j-1}}) & \text{for } s,t \in (\tau_{j-1}, \tau_j] \\ K^*(\frac{s-\tau_j}{1-\tau_j}, \frac{t-\tau_j}{1-\tau_j}, \frac{1-\tau_{j-1}}{1-\tau_{j-1}}) & \text{for } s,t \in (\tau_j, 1] \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

5 Asymptotic properties of LACOSSO.

Let the L_2 norm of a function evaluated at the data points be denoted

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f^2(x_i)$$

The following theorem states that LACOSSO attains the optimal convergence rate for nonparametric regression estimators. The proof is provided in Section A.3 of the Supplementary Material.

Theorem. Consider the regression model $y_i = f_0(x_i) + \epsilon_i,\ i = 1, 2, ..., n$, where $x_i's$ are given values of a covariate in $[0, 1]$, and $\epsilon_i's$ are independent $N(0, \sigma^2)$ errors. Assume f_0 lies in S^2 with S^2 being the second order Sobolev space.

Let $\hat{f}$ be defined as in (8) with $w_j = 1$ for all j and let $I(f) = \int_0^1 [f''(x)]^2 dx$. Then (i) if f_0 is a nonlinear function, and $\lambda_n^{-1} = O_p(n^{2/5}) I^{3/10}(f_0)$, then $\|\hat{f} - f_0\|_n = O_p(\lambda_n) I^{1/2}(f_0)$; (ii) if f_0 is a linear function, then $\|\hat{f} - f_0\|_n = O_p(\max(n\lambda_n)^{-2/3}, n^{-1/2})$.

Remark 1. if $\lambda_n \sim n^{-2/5}$ then $\|\hat{f} - f_0\|_n = O_p(n^{-2/5})$ which is optimal for nonparametric regression estimators.

Remark 2. Here it is assumed that $w_j = 1$, but this could be relaxed. All that is really needed is for $w_j = O_p(1)$ and $w_j^{-1} = O_p(1)$ in order for the proof to go through.

6 Example Results

In this section the performance of LACOSSO is evaluated on several simulated data sets. The results are compared to those from several other competing methods. The methods included in these simulations are:

LOCO − The Loco-Spline procedure with tuning parameter selection via 5-fold CV as described in Storlie et al. (2010).

SAS(5) − the version of the spatially adaptive smoothing spline suggested in Pintore et al. (2006) which uses piecewise constant (with 5 bins since this had the best performance in their paper) for $\lambda(x)$.

TRAD − the traditional smoothing spline (TRAD) with tuning parameter chosen via GCV.

LOKERN − local kernel regression with plug-in local bandwidth as provided by the R package lokern. This procedure uses a second order kernel with a plug-in estimate of the asymptotically optimal local bandwidth.

MARS − Multivariate Adaptive Regression Splines (Friedman 1991) as provided by the R package polymars. This procedure uses regression splines with spatially adaptive knot placement.

LACOSSO($\boldsymbol{\gamma}$,$\boldsymbol{p}$) − Locally Adaptive COSSO procedure with tuning parameter selection via GCV with weight power γ and p bins ($\tau_j's$ placed at evenly spaced quantiles of x). In the simulations results are reported for all combinations of $\gamma = 0, 1$, and $p = 5, 10, 20$.

6.1 Mexican Hat Function

The first test problem which we call the Mexican hat function is a quadratic function with a sharp Gaussian bump in the middle of the domain. Specifically the function is given by

$$f(x) = -1 + 1.5x + 0.02\phi_{0.02}(x - 0.5) \tag{24}$$

where $\phi_\sigma(x-\mu)$ is the $N(\mu, \sigma^2)$ density evaluated at x. A simple random sample of size n is generated from from $x_i \sim \text{Unif}\,(0,1)$, $i = 1, 2, ..., n$. Then $Y_i = f(x_i) + \epsilon_i$, is generated, where $\epsilon_i \sim N(0, 0.25)$.

Figure 1 displays the data along with the corresponding fits from LACOSSO and traditional smoothing spline for a typical realization with $n = 100$. Here it can be seen that LACOSSO-spline is able to both better capture the peak and stay smooth where the function is flat. On the other hand, see how the traditional smoothing spline "chases" data points in areas where the true function is flat.

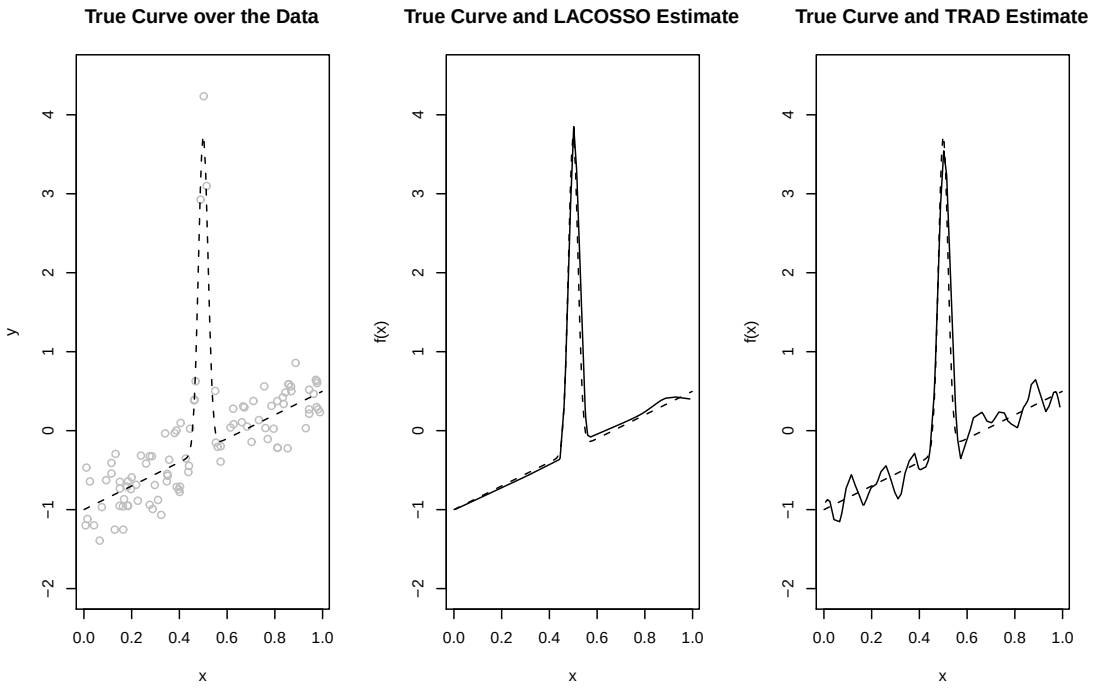


Figure 1: Mexican hat function. Left: Data generated from the Mexican hat Function with n=100 along with the true function. Middle: The LACOSSO (1,20) estimate (solid) with true function (dashed). Right: The traditional smoothing spline estimate (solid) with the true function (dashed)

.

Tier one of Table 1 compares the performance on the Mexican hat example for these methods as sample size increases. The reported summary statistics are the average mean squared error (AMSE) and the percent best. The AMSE is the average of the MSE over

100 realizations at the respective sample sizes. Here we are using the definition of MSE which averages squared errors at the data points, i.e., $MSE = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2$. The percent best is the percentage of the 100 realizations that a given method had the smallest MSE among the competing methods.

From Table 1 it is clear that LACOSSO has a very competitive performance for all sample sizes on the Mexican hat example example. LACOSSO (1,20) had the smallest MSE for approximately 40% of the realizations for each sample size.

6.2 Dampened Harmonic Motion

The next problem is a dampened harmonic motion also known as the spring equation. Functions with this type of behavior are common to just about any structural engineering problem. The spring equation is given by

$$f(x) = a\ \exp\{-b(1-x)\}\cos\{w(1-x)\} \tag{25}$$

The parameter values of $a = 1, b = 30, w = 30\pi$ have been chosen to produce the data for this simulation. Again, $x_i \sim \text{Unif}\,(0,1)$, $i = 1, 2, ..., n$ with $Y_i = f(x_i) + \epsilon_i$, but here $\epsilon \sim N(0, 0.05)$.

Figure 2 displays the data and the corresponding fits from LACOSSO and traditional smoothing spline for a typical realization with $n = 100$. Here it can be seen that the LACOSSO-spline better captures the behavior of this function. The traditional smoothing estimate does not capture the higher amplitude oscillation as well as LACOSSO does and, again, allows for the undesirable behavior of "chasing" points in areas where the true function is flat.

Tier two of Table 1 summarizes the performance on the dampened harmonic example for sample sizes $n = 100, 200,$ and 300. In this example, the proposed method has a performance as good as the one shown by LOCO and SAS(5). However, LACOSSO

Mexican Hat			
	$n = 100$	$n = 200$	$n = 300$
LOCO-SPLINE	158.40 (5.82)	52.69 (2.71)	37.32 (1.88)
SAS(5)	106.47 (5.40)	55.10 (3.09)	35.45 (1.59)
TRAD	205.27 (4.62)	116.96 (2.62)	81.77 (1.57)
LOKERN	342.03 (20.8)	157.31 (6.12)	108.64 (3.05)
MARS	655.64 (40.3)	645.70 (44.5)	493.14 (33.1)
LACOSSO (0,5)	173.38 (5.74)	89.93 (3.07)	61.78 (2.04)
LACOSSO (1,5)	103.56 (4.39)	48.66 (1.79)	31.55 (1.29)
LACOSSO (0,10)	145.07 (5.30)	77.20 (2.79)	52.98 (1.80)
LACOSSO (1,10)	88.59 (3.87)	44.81 (1.78)	30.62 (1.21)
LACOSSO (0,20)	134.87 (5.40)	74.61 (2.84)	46.60 (1.64)
LACOSSO (1,20)	85.08 (4.00)	43.67 (1.99)	27.91 (1.16)
Dampened Harmonic			
	$n = 100$	$n = 200$	$n = 300$
LOCO-SPLINE	3.75 (0.40)	2.99 (0.29)	2.33 (0.08)
SAS(5)	3.33 (0.17)	1.92 (0.07)	1.30 (0.04)
TRAD	9.40 (0.26)	5.86 (0.11)	4.09 (0.06)
LOKERN	31.50 (2.15)	22.98 (1.50)	19.68 (1.55)
MARS	44.13 (3.17)	52.02 (2.44)	68.37 (1.97)
LACOSSO (0,5)	5.12 (0.19)	2.91 (0.09)	1.89 (0.06)
LACOSSO (1,5)	3.59 (0.16)	2.06 (0.08)	1.32 (0.04)
LACOSSO (0,10)	3.77 (0.16)	2.11 (0.08)	1.37 (0.04)
LACOSSO (1,10)	3.17 (0.16)	1.82 (0.09)	1.23 (0.05)
LACOSSO (0,20)	3.67 (0.15)	2.03 (0.08)	1.31 (0.04)
LACOSSO (1,20)	3.31 (0.17)	2.04 (0.09)	1.37 (0.04)
Rapid Change			
	$n = 100$	$n = 200$	$n = 300$
LOCO-SPLINE	3.54 (0.22)	1.61 (0.09)	1.13 (0.05)
SAS(5)	4.15 (0.25)	2.05 (0.10)	1.36 (0.05)
TRAD	5.49 (0.16)	3.05 (0.07)	2.15 (0.04)
LOKERN	7.41 (0.32)	3.76 (0.12)	2.68 (0.07)
MARS	5.34 (0.34)	3.37 (0.20)	2.61 (0.10)
LACOSSO (0,5)	4.26 (0.11)	2.43 (0.07)	1.81 (0.06)
LACOSSO (1,5)	3.16 (0.12)	1.87 (0.08)	1.33 (0.05)
LACOSSO (0,10)	4.32 (0.12)	2.46 (0.07)	1.86 (0.05)
LACOSSO (1,10)	2.69 (0.13)	1.55 (0.09)	1.07 (0.04)
LACOSSO (0,20)	4.83 (0.12)	2.53 (0.07)	1.87 (0.05)
LACOSSO (1,20)	2.77 (0.14)	1.45 (0.09)	0.98 (0.04)

Table 1: Table 1: Results of 100 Realizations from Mexican hat, Dampened Harmonic, and Rapid Change examples. The the mean square error averaged over the 100 realizations (AMSE) with standard error in parentheses is presented for sample sizes of $n = 100, 200, 300$.

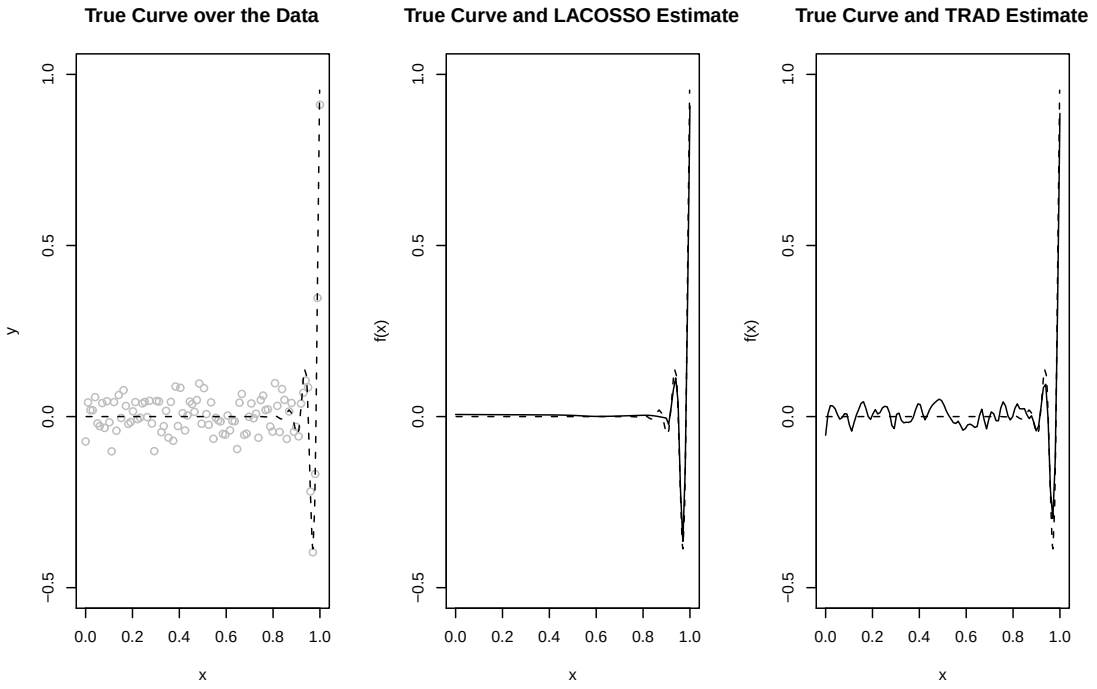


Figure 2: Dampened harmonic function. Left: Data generated from the dampened harmonic function with n=100 along with the true function. Middle: The LACOSSO (1,10) estimate (solid) with true function (dashed). Right: The traditional smoothing spline estimate (solid) with the true function (dashed)

.

(1,10) has smaller MSE in, roughly, 30% of the realizations for all sample sizes, almost twice as much as SAS(5), its closest competitor.

6.3 Rapid Change Function

The rapid change function is defined as

$$f(x) = \frac{0.8}{1 + \exp[-75(x - 0.8)]} \tag{26}$$

Once again, $x_i \sim \text{Unif}\,(0, 1)$ with $Y_i = f(x_i) + \epsilon_i$ and $\epsilon_i \sim N(0, 0.05)$.

Figure 3 displays the data and the corresponding fits from the traditional smoothing spline for a typical realization with $n = 100$. The smoothing spline estimate is very rough

overall whereas the LACOSSO-spline is able to fit the true function just as well in the rapid change region as in the other regions.

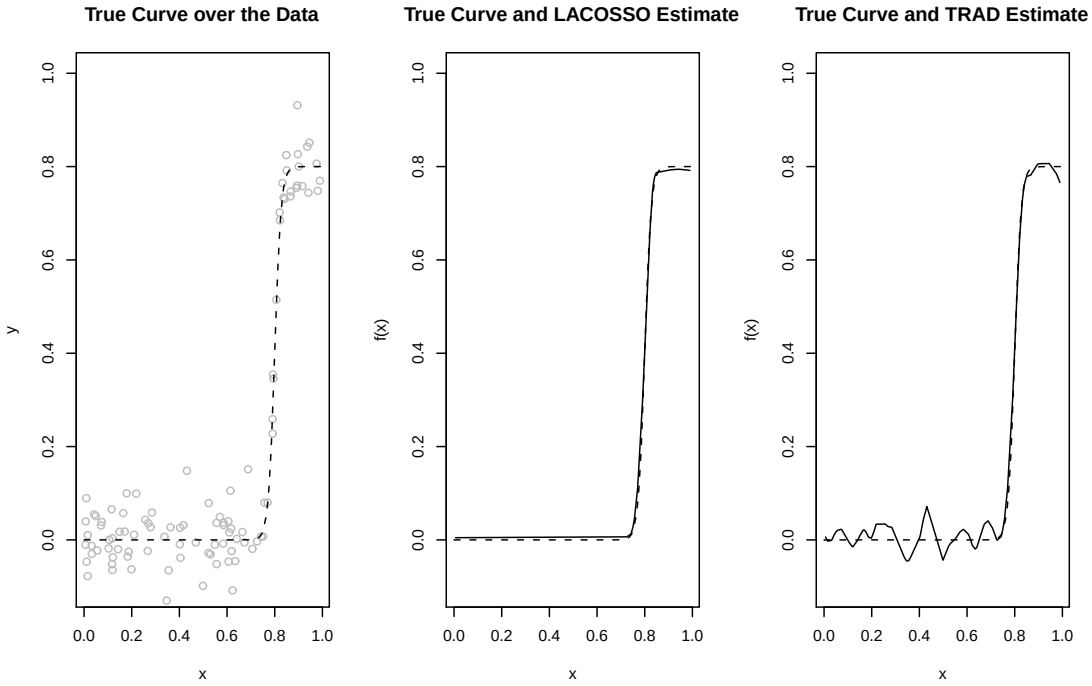


Figure 3: Rapid change function. Left: Data generated from the rapid change function with n=100 along with the true function. Middle: The LACOSSO (1,20) estimate (solid) with true function (dashed). Right: The traditional smoothing spline estimate (solid) with the true function (dashed)

.

Tier three of Table 1 summarizes the results of the simulations from this example. In this example, LACOSSO (1, 10) and (1, 20) have a lower AMSE than the other methods at all sample sizes. The only method that compares to these two is LOCO-Spline.

6.4 Waste Isolation Pilot Plant Example

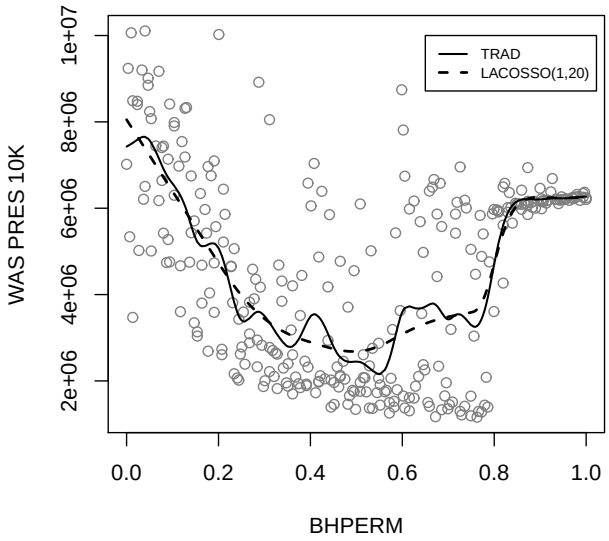
Nonparametric regression techniques have become a popular tool for analyzing complex computer model output (Storlie & Helton 2008, Reich, Storlie & Bondell 2009, Storlie, Swiler, Helton & Sallaberry 2009). Here we consider a two-phase fluid flow simulation

study (Vaughn, Bean, Helton, Lord, MacKinnon & Schreiber 2000) carried out by Sandia National Laboratories as part of the 1996 compliance certification application for the Waste Isolation Pilot Plant (WIPP) in New Mexico. The computer model simulates the waste panel's condition 10,000 years after the waste panel has been penetrated by a drilling intrusion.

A small example from these results is considered in this presentation. In particular, the modeling case corresponding to a drilling intrusion at 1,000 yr that penetrates both the WIPP repository and an underlying region of pressurized brine is used as an example (i.e., an E1 intrusion at 1,000 yr in the terminology of the 1996 WIPP CCA; see Table 6, (Helton, Martell & Tierney 2000)). Specifically, we investigate the relationship between waste pressure at 10,000 years (*WAS_PRES.10K*) and Bore Hole Permeability (*BHPERM*). This is a very interesting example, because at large values of BHPERM, so much brine flows down the borehole that the repository saturates and rises to hydrostatic pressure. This phenomenon can be seen in Figure 4 where there is an abrupt change in *WAS_PRES.10K* around *BHPERM*= 0.8.

Figure 4 displays the fit of the traditional smoothing spline along with the LACOSSO(1,10) estimate, both using 5-fold CV for tuning parameter selection. This example benefits from the local approach to smoothing as LACOSSO clearly results in a much more appealing main effect estimate for *BHPERM*. The LACOSSO estimate is a smooth function for *BHPERM* < 0.8, makes an abrupt change at *BHPERM* $\approx$ 0.8, then smooth

Figure 4: Scatterplot of *WAS_PRES.10K* versus *BHPERM* along with the LACOSSO fit (dashed) and the Traditional Smoothing Spline fit (solid).



again. The traditional smoothing spline on the other hand has to select a small tuning parameter to reduce the penalty in order to adequately capture the change at *BHPERM* ≈ 0.8. This results in "chasing" data points and an undesirable rough behavior of the estimated function for $BHPERM < 0.8$

6.5 Computational Time

The CPU times required to fit each of the models are displayed in Figure 5 for each of the methods (lokern, trad, mars, sas(5), lacosso(1,5), lacosso(1,10), lacosso(1,20), loco-spline), for the Mexican Hat function presented in Section 6.1. Computation times on the other examples from Sections 6.2 and 6.3 were similar. All models were fit using a commodity machine with Pentium quad core 2.66 GH processors (although parallel computing was not used). The times presented in Figure 5 represent average computation time over 10 different datasets.

To make direct computational time comparisons, all methods are fully implemented in R. Substantial computational savings would be expected if they were fully or even partially implemented in a compiled language such as C. This is particularly true for the Loco-spline procedure as it requires a numerical integration to evaluate each element of the Gram matrix.

7 Conclusions

A new nonparametric regression estimator, LACOSSO, is obtained via solving a regularization problem with a novel adaptive penalty on the sum of functional norms which allows for a locally varying smoothness of the resulting estimate. The effectiveness of this approach as a scatterplot smoother is demonstrated when compared to the traditional smoothing spline and other more spatially adaptive methods. LACOSSO machinery can be effectively transferred into higher dimensional problems and non-continuous responses

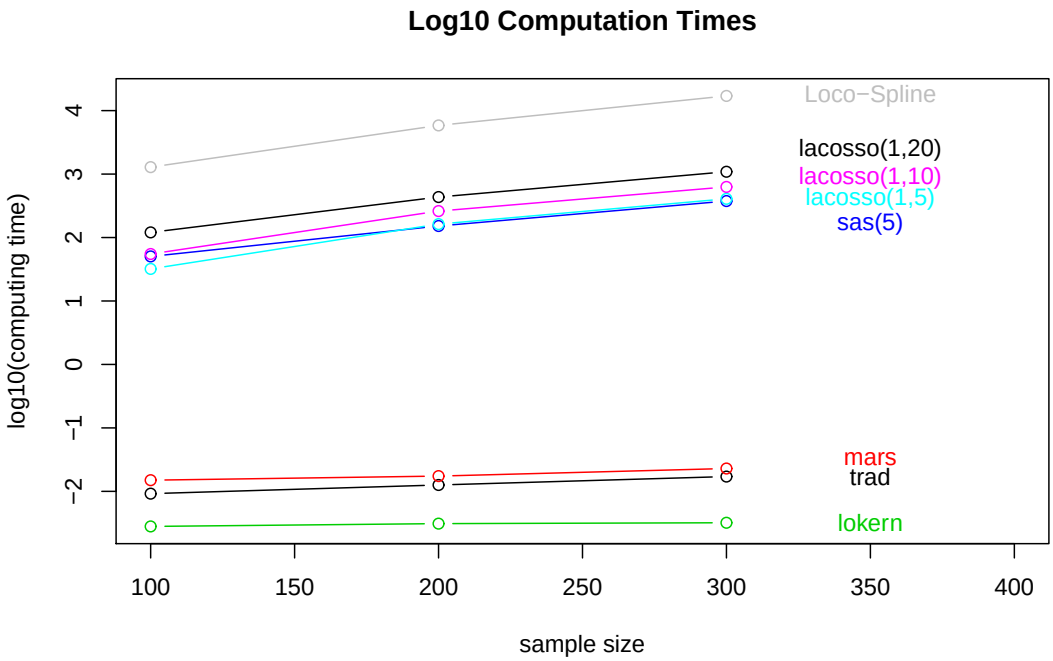


Figure 5: log10 (average computing times) versus sample size for the Mexican hat example

.

(Bernoulli data, Poisson data, etc.). Its performance is not adversely affected when allowing for more flexibility (more bins) unlike other methods that have a tendency to overfit. In fact the performance of LACOSSO seems to improve with the addition of bins, which is a contrast to the method of Pintore et al. (2006), for example, with 10 and 20 bins. This behavior can be attributed to the formulation of the minimization in the COSSO like framework, which involves only one tuning parameter, instead of one tuning parameter per bin. The MSE asymptotic optimality of this method has also been established.

References

Berlinet, A. & Thomas-Agnan, C. (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Norwell, MA: Kluwer Academic Publishers.

Cox, D. (1983), 'Asymptotics for m-type smoothing splines', *Annals of Statistics* **11**(2), 530–551.

Eubank, R. (1999), *Nonparametric Regression and Spline Smoothing*, CRC Press.

Friedman, J. (1991), 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics* **19**, 1–141.

Friedman, J. & Silverman, B. (1989), 'Flexible parsimonious smoothing and additive modeling (with discussion)', *Technometrics* **31**, 3–39.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer-Verlag, New York, NY.

Hansen, M. & Kooperberg, C. (2002), 'Spline adaptation in extended linear models (with discussion)', *Statistical Science* **17**, 2–51.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC.

Helton, J., Martell, M. & Tierney, M. (2000), 'Characterization of subjective uncertainty in the 1996 performance assessment for the Waste Isolation Pilot Plant', *Reliability Engineering and System Safety* **69**(1-3), 191–204.

Lin, Y. & Zhang, H. (2006), 'Component selection and smoothing in smoothing spline analysis of variance models', *Annals of Statistics* **34**(5), 2272–2297.

Luo, Z. & Wahba, G. (1997), 'Hybrid adaptive splines', *Journal of the American Statistical Association* **92**(437), 107–116.

Nosedal-Sanchez, A., Storlie, C., Lee, T. & Christensen, R. (2011), 'Reproducing kernel hilbert spaces for penalized regression: A tutorial', *American Statistician* (in review).

Parzen, E. (1961), 'An approach to time series analysis', *Annals of Statistics* **32**, 951–989.

Parzen, E. (1962), *Stochastic Processes*, 1st edition edn, Holden-Day.

Pintore, A., Speckman, P. & Holmes, C. (2006), 'Spatially adaptive smoothing splines', *Biometrika* **93**(1), 113–125.

Reich, B., Storlie, C. & Bondell, H. (2009), 'Variable selection in bayesian smoothing spline anova models: Application to deterministic computer codes', *Technometrics* **51**(2), 110–120.

Schimek, M., ed. (2000), *Smoothing and Regression: Approaches, Computation, and Application*, John Wiley & Sons, Inc., New York, NY.

Speckman, P. (1985), 'Spline smoothing and optimal rates of convergence in nonparametric regression-models', *Annals of Statistics* **13**(3), 970–983.

Stone, C., Hansen, M., Kooperberg, C. & Truong, Y. (1997), '1994 wald memorial lectures - polynomial splines and their tensor products in extended linear modeling', *Annals of Statistics* **25**(4), 1371–1425.

Storlie, C., Bondell, H. & Reich, B. (2010), 'A locally adaptive penalty for estimation of functions with varying roughness', *Journal of Computational and Graphical Statistics* **19**(3), 569–589.

Storlie, C., Bondell, H., Reich, B. & Zhang, H. (2011), 'Surface estimation, variable selection, and the nonparametric oracle property', *Statistica Sinica* **21**(2), 679–705.

Storlie, C. & Helton, J. (2008), 'Multiple predictor smoothing methods for sensitivity analysis: Description of techniques', *Reliability Engineering and System Safety* **93**(1), 28–54.

Storlie, C., Swiler, L., Helton, J. & Sallaberry, C. (2009), 'Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models', *Reliability Engineering and System Safety* **94**, 1735–1763.

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press.

Vaughn, P., Bean, J., Helton, J., Lord, M., MacKinnon, R. & Schreiber, J. (2000), 'Representation of two-phase flow in the vicinity of the repository in the 1996 performance assessment for the Waste Isolation Pilot Plant', *Reliability Engineering and System Safety* **69**(1-3), 205–226.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline anova for exponential families, with application to the WESDR', *Annals of Statistics* **23**, 1865–1895.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

A Supplementary Material

A.1 Calculations to find the r.k.

Here we give the specific calculations of the covariance functions K_1 and K_2 for Section 4.2. We then demonstrate that K_1 and K_2 are the r.k.'s for S_1^* and S_2^*, respectively.

Case 1. $0 \leq s \leq t \leq \tau_1$

First, we need to recall that the mean value and covariance of a Wiener Process are given by

$$m(t) = E[X(t)] = 0 \tag{A1}$$

$$Cov[X(s), X(t)] = \min(s, t) \tag{A2}$$

Now, note that the expectation of $Z_1(t)$ is equal to zero.

$$E\left[\int_0^t X(\nu)d\nu\right] = \int_0^t E[X(\nu)]d\nu = 0 \tag{A3}$$

This implies that the covariance between $Z_1(s)$ and $Z_1(t)$ is given by

$$\begin{aligned}
Cov[Z_1(s)Z_1(t)] &= E\left[Z_1(s)Z_1(t)\right] \\
&= E\left[\int_0^s X(y)dy \int_0^t X(u)du\right] \\
&= E\left[\int_0^s \int_0^t X(y)X(u)dydu\right] \\
&= \int_0^s \int_0^t E\left[X(y)X(u)\right] dydu \\
&= \int_0^s \int_0^t \min(y, u)dydu \\
&= \int_0^s \left(\int_0^u ydy + \int_u^t udy\right) du \\
&= s^2\left(\frac{t}{2} - \frac{s}{6}\right) = \frac{s^2 t}{2} - \frac{s^3}{6}
\end{aligned}$$

or in general for $0 < s < \tau_1$, $0 < t < \tau_1$,

$$E\left[Z_1(s)Z_1(t)\right] = \frac{\min^2(s,t)\max(s,t)}{2} - \frac{\min^3(s,t)}{6} \tag{A4}$$

Case 2. If t and s are in $[\tau_1, 1]$ and $s < t$.

First, note that $E(Z_1(t)) = 0$.

To determine the covariance between $Z_1(s)$ and $Z_1(t)$ we need to find its product

$$Z_1(s)Z_1(t) = [Z_1(\tau_1) + (s-\tau_1)X(\tau_1)][Z_1(\tau_1) + (t-\tau_1)X(\tau_1)]$$

$$= Z_1^2(\tau_1) + Z_1(\tau_1)X(\tau_1)(t-\tau_1) + Z_1(\tau_1)X(\tau_1)(s-\tau_1) + (s-\tau_1)(t-\tau_1)X^2(\tau_1)$$

$$= Z_1^2(\tau_1) + Z_1(\tau_1)X(\tau_1)[(t-\tau_1) + (s-\tau_1)] + (s-\tau_1)(t-\tau_1)X^2(\tau_1)$$

Now,

$$E[Z_1(s)Z_1(t)] = E[Z_1^2(\tau_1)] + [(t-\tau_1) + (s-\tau_1)]E[Z_1(\tau_1)X(\tau_1)] + (s-\tau_1)(t-\tau_1)E[X^2(\tau_1)] \tag{A5}$$

We know, from case 1, that

$$E[Z_1^2(\tau_1)] = \frac{\tau_1^3}{2} - \frac{\tau_1^3}{6} = \frac{\tau_1^3}{3} \tag{A6}$$

$$E[X(\tau_1)Z_1(\tau_1)] = E\left[\int_0^{\tau_1} X(y)X(\tau_1)dy\right] \tag{A7}$$

$$= \int_0^{\tau_1} \min(y, \tau_1)dy = \frac{\tau_1^2}{2} \tag{A8}$$

$$E[X^2(\tau_1)] = \min(\tau_1, \tau_1) = \tau_1 \tag{A9}$$

Then, substituting (A6), (A8) and (A9) into (A5) gives

$$E[Z_1(s)Z_1(t)] = \frac{\tau_1^3}{3} + [(t-\tau_1) + (s-\tau_1)]\left[\frac{\tau_1^2}{2}\right] + (s-\tau_1)(t-\tau_1)\tau_1 \tag{A10}$$

Case 3. $0 \le s \le \tau_1 \le t \le 1$.

From the two cases discussed above, we have that $E(Z_1(t)) = 0$. To find the covariance

of $Z_1(t)$, first we need the product $Z_1(s)Z_1(t)$

$$\begin{aligned} Z_1(s)Z_1(t) &= \left[\int_0^s X(\nu)d\nu\right]\left[\int_0^{\tau_1} X(\nu)d\nu + (t-\tau_1)X(\tau_1)\right] \\ &= \left[\int_0^s X(\nu)d\nu\right]\left[\int_0^s X(\nu)d\nu + \int_s^{\tau_1} X(\nu)d\nu + (t-\tau_1)X(\tau_1)\right] \\ &= \left[\int_0^s X(\nu)d\nu\right]^2 + \left[\int_0^s X(\nu)d\nu\right]\left[\int_s^{\tau_1} X(\nu)d\nu\right] + (t-\tau_1)X(\tau_1)\left[\int_0^s X(\nu)d\nu\right] \\ &= [Z_1(s)]^2 + [Z_1(s)]\,[Z_1(\tau_1) - Z_1(s)] + (t-\tau_1)X(\tau_1)\,[Z_1(s)] \end{aligned}$$

Now,

$$E[Z_1(s)Z_1(t)] = E\,[Z_1(s)]^2 + E\,[Z_1(s)]\,[Z_1(\tau_1) - Z_1(s)] + (t-\tau_1)E\,[X(\tau_1)Z_1(s)] \quad \text{(A11)}$$

From calculations in cases 1 and 2, we know that

$$E[Z_1^2(s)] = Var[Z_1(s)] = \frac{s^3}{3} \quad \text{(A12)}$$

$$E[X(\tau_1)Z_1(s)] = \frac{s^2}{2} \quad \text{(A13)}$$

We also need to find the following expectation

$$E\,[Z_1(s)]\,[Z_1(\tau_1) - Z_1(s)] = Cov\,[Z_1(s)Z_1(\tau_1)] - Var\,[Z_1(s)] \quad \text{(A14)}$$

$$= \left[\frac{s^2\tau_1}{2} - \frac{s^3}{6} - \frac{s^3}{3}\right] = \left[\frac{s^2\tau_1 - s^3}{2}\right] \quad \text{(A15)}$$

Finally, substituting (A12),(A13) and (A15) into (A11) we have that

$$E[Z_1(s)Z_1(t)] = \frac{s^3}{3} + \left[\frac{\tau_1 s^2 - s^3}{2}\right] + \frac{(t-\tau_1)s^2}{2} = \frac{ts^2}{2} - \frac{s^3}{6} \quad \text{(A16)}$$

or for general case $0 \le s \le \tau_1 \le t \le 1$ or $0 \le t \le \tau_1 \le s \le 1$

$$E[Z_1(s)Z_1(t)] = \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6} \quad \text{(A17)}$$

A.2 Proving that $R_1 = K_1$ is the r.k. of S_1^*.

We now prove that $R_1^* = K_1$ has the "reproducing property" and hence is the r.k. of the space S_1^*. We split this into two cases: (1) $t \in [0, \tau_1]$ and (2) $t \in [\tau_1, 1]$.

Case 1. Assume that $t \in [0, \tau_1]$, by definition the inner product between $f(\cdot)$ and $K(\cdot, t)$ is given by

$$\langle f(t), K(\cdot, t)\rangle = \int_0^1 \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds$$
$$= \int_0^t \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds + \int_t^{\tau_1} \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds + \int_{\tau_1}^1 \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds$$

Note that the second term and the third term on the right hand side of the last equation are equal to zero. One can see this by using our definition of $K(s,t)$, case 1 in (A4) and case 3 in (A17) respectively, and finding the second partial derivative of $K(s,t)$ for each case with respect to s when $s > t$.

Doing this we have that

$$\langle f(t), K(s,t)\rangle = \int_0^t \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds = \int_0^t (t-s) f''(s) ds. \tag{A18}$$

Integrating by parts this last expression

$$\langle f(t), K(s,t)\rangle = (t-t) f'(t) - (t-0) f'(0) + f(t) - f(0). \tag{A19}$$

Finally, recall that $f'(0) = f(0) = 0$ which implies that

$$\langle f(t), K(s,t)\rangle = f(t). \tag{A20}$$

Case 2. Assume that $t \in [\tau_1, 1]$, by definition the inner product between $f(\cdot)$ and $K(\cdot, t)$ is given by

$$\langle f(t), K(s,t)\rangle = \int_0^1 \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds$$
$$= \int_0^{\tau_1} \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds + \int_{\tau_1}^t \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds + \int_t^1 \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds$$

Note that the second term and third terms on the right hand side of the equation shown above are equal to zero, by our definition of $K(s,t)$, so that

$$\langle f(t), K(s,t)\rangle = \int_0^{\tau_1} \frac{\partial^2}{\partial s^2} K(s,t) f''(s) ds = \int_0^{\tau_1} (t-s) f''(s) ds \tag{A21}$$

$$= \int_0^{\tau_1} (t-s) f''(s) ds = (t-\tau_1) f'(\tau_1) - t f'(0) + f(\tau_1) - f(0) \tag{A22}$$

Recalling that $f'(0) = f(0) = 0$ we have that

$$\langle f(t), K(s,t)\rangle = f(\tau_1) + (t-\tau_1) f'(\tau_1) = f(t) \tag{A23}$$

Applying the same arguments to the shifted and rescaled versions of s and t, one can prove that $R_2^* = K_2$ is the r.k. for S_2^* as well.

A.3 Proof of the Convergence Theorem.

Before presenting a proof for the result regarding optimal MSE convergence, we state a definition and a lemma necessary for the proof.

Definition. (Entropy for the supremum norm) For a function space G, let $N_\infty(\delta, G)$ be the smallest value of N such that there exists $\{g_j\}_{j=1}^N$ with

$$\sup_{g\in G} \min_{j=1,..,N} |g - g_j|_\infty \leq \delta.$$

Then $H_\infty(\delta, G) = \log N_\infty(\delta, G)$ is called the δ-entropy of G for the supremum norm. Where $|g|_\infty = \sup_{x\in X} |g(x)|$.

Lemma 1. Consider a regression model $y_i = g_0(x_i) + \epsilon_i$, $i = 1, ..., n$ where g_0 is known to lie in a class G of functions, x_i are given covariates in $[0,1]^p$, and ϵ_i are independent $N(0, \sigma^2)$ errors. Let $I : G \to [0, \infty)$ be a pseudo-norm on G. Define

$$\hat{g} = \arg \min_{g\in G} \frac{1}{n} \sum_{i=1}^n \{y_i - g(x_i)\}^2 + \lambda_n^2 I(g)$$

Assume

$$H_\infty\left(\delta, \{\frac{g-g_0}{I(g)+I(g_0)} : g \in G, I(g)+I(g_0)>0\}\right) \leq A\delta^{-\alpha} \tag{A24}$$

for all $\delta > 0$, $n \geq 1$ and some $A > 0$, $0 < \alpha < 2$. Here H_∞ stands for the entropy for the supreme norm. Then i) if $I(g_0) > 0$ and $\lambda_n^{-1} = O_p(n^{\frac{1}{2+\alpha}})I^{\frac{2-\alpha}{4+2\alpha}}(g_0)$, we have $\|\hat{g} - g_0\|_n = O_p(\lambda_n)I^{1/2}(g_0)$; ii) if $I(g_0) = 0$ we have $\|\hat{g} - g_0\|_n = O_p(n^{\frac{-1}{2-\alpha}})\lambda_n^{\frac{-2\alpha}{2-\alpha}}$

Before actually proving the theorem we give a brief sketch of the proof to provide more clarity. First, note that if we prove that G, the class of functions we are interested in, is bounded in entropy then we can use Lemma 1 and we are done. It turns out that with the supremum norm we have problems with the linear part of our functions. That is, Lemma 1 cannot be used directly since (A24) is not satisfied in our case. To see this define the following set of functions

$$F = \{f(x) = \alpha + \beta x,\ x \in [0,1],\ \alpha, \beta \in \Re\} \tag{A25}$$

Note that given any $\delta > 0$ and any finite set of functions g_j's we can find a function $f \in F$ such that $|f - g_j|_\infty > \delta$. Therefore, we decompose our functional space into two parts: linear part and non-linear part. Then we deal with each of these components separately. After we give a rate of convergence for the linear part, we deal with the non-linear part. We will show that the entropy of the functional space of the nonlinear part has the form $A^*\delta^{1/2}$, for some A^* and the desired result follows from Lemma 1.

Proof of the Theorem. Lemma 1 cannot be used directly since (A24) is not satisfied in our case. Therefore, to apply lemma 1 we have to decompose the space of functions in two parts: linear part and non-linear part. This problem can be dealt with with the following arguments. For any $f \in S^2$, we can write

$$f(x) = b_0 + b_1 x + f_1(x) + ... + f_p(x) = g_1(x) + g_2(x)$$

where $g_1(x) = b_0 + b_1 x$, $g_2(x) = f_1(x) + ... + f_p(x)$, $f_j \in S_j^*$, $\sum_{i=1}^n f_j(x_i) = 0$ and

$\sum_{i=1}^{n} x_i f_j(x_i) = 0$ for $j = 1, 2, ..., p$.

Similarly, for the unknown underlying function f_0, write

$$f_0(x) = b_{00} + b_{01}x + f_{01}(x) + ... + f_{0p}(x) = g_{01}(x) + g_{02}(x)$$

where $g_{01}(x) = b_{00} + b_{01}x$, $g_{02}(x) = f_{01}(x) + ... + f_{0p}(x)$, $f_{0j} \in S_j^*$, $\sum_{i=1}^{n} f_{0j}(x_i) = 0$ and $\sum_{i=1}^{n} x_i f_{0j}(x_i) = 0$ for $j = 1, 2, ..., p$. Then, by construction $\sum_{i=1}^{n}\{g_{01}(x_i) - g_1(x_i)\}\{g_{02}(x_i) - g_2(x_i)\} = 0$.

Then we can write $\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda_n J(f)$, where
$J(f) = \sum_{j=1}^{p}\left\{\int_{\tau_{j-1}}^{\tau_j}[f''(x)]^2 dx\right\}^{1/2}$, as

$$\frac{1}{n}\sum_{i=1}^{n}\{(g_{01}(x_i) - g_1(x_i)) + (g_{02}(x_i) - g_2(x_i) + \epsilon_i)\}^2 + \lambda_n J(g)$$

$\frac{1}{n}\sum_{i=1}^{n}\{(g_{01}(x_i) - g_1(x_i))\}^2 + \frac{2}{n}\sum_{i=1}^{n}(g_{01}(x_i) - g_1(x_i))(g_{02}(x_i) - g_2(x_i) + \epsilon_i)$
$+ \sum_{i=1}^{n}(g_{02}(x_i) - g_2(x_i) + \epsilon_i)^2 + \lambda_n J(g)$.

Due to the conditions imposed above (we have those conditions to guarantee that g_1 and g_2 are orthogonal under the empirical inner product), we have

$$\frac{1}{n}\sum_{i=1}^{n}\{(g_{01}(x_i) - g_1(x_i))\}^2 + \frac{2}{n}\sum_{i=1}^{n}(g_{01}(x_i) - g_1(x_i))\epsilon_i + \sum_{i=1}^{n}(g_{02}(x_i) - g_2(x_i) + \epsilon_i)^2$$
$$+ \lambda_n J(g_2)$$

Therefore the corresponding g_1 to the f which minimizes (8) must minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{(g_{01}(x_i) - g_1(x_i))\}^2 + \frac{2}{n}\sum_{i=1}^{n}(g_{01}(x_i) - g_1(x_i))\epsilon_i.$$

By example 9.3.1 of van de Geer (2000), page 152, we have that $\hat{g}_1$ converges with rate $n^{-1/2}$.

On the other hand, the non-linear part, $\hat{g}_2$ must minimize

$$\frac{1}{n}\sum_{i=1}^{n}[g_{02}(x_i) - g_2(x_i)]^2 + \lambda_n J(g_2)$$

Let $G = \{g \in S^2 : g(x) = f_1(x) + ... + f_p(x)$ with $f_j \in S_j^*$, $\sum_{i=1}^n f_j(x_i) = 0$, and $\sum_{i=1}^n x_i f_j(x_i) = 0$, $j = 1, 2, ..., p\}$.

We can now apply lemma 1 with $I = J$ and $\alpha = 1/2$. All that remains to be shown is that (A24) is satisfied. The conclusion of the Theorem then follows from the conclusion of lemma 1.

Let $J^*(g) = \int_0^1 [f''(x)]^2 dx$. From page 168 of van de Geer (2000), note that

$$H_\infty(\delta, \{g \in G : J^*(g) \leq 1\}) \leq A\delta^{-1/2}.$$

Also,

$$J^*(g) = \int_0^1 [f''(x)]^2 dx \leq \left(\sum_{j=1}^p \left\{ \int_{\tau_{j-1}}^{\tau_j} [f''(x)]^2 dx \right\}^{1/2} \right)^2 = J^2(g)$$

Thus $J(g) \leq 1$ implies that $J^*(g) \leq 1$ so that $\{g \in G : J(g) \leq 1\} \subset \{g \in G : J^*(g) \leq 1\}$. Now if $\{g \in G : J^*(g) \leq 1\}$ can be covered by N balls of radius δ, then $\{g \in G : J(g) \leq 1\}$ can be covered by the same balls since it is a smaller set. Hence,

$$H_\infty(\delta, \{g \in G : J(g) \leq 1\}) \leq A\delta^{-1/2}.$$

Lastly, noting that $J(g - g_0) \leq J(g) + J(g_0)$ for any $g \in G$, we see that (A24) is satisfied. The conclusion of the Theorem then follows from the conclusion of lemma 1.